A Task-based Evaluation of French Morphological Resources and Tools

Delphine Bernhard, Bruno Cartoni, Delphine Tribout

Linguistic **I**ssues in **L**anguage **T**echnology – **LiLT**
Submitted, October 2011

A Task-Based Evaluation of French Morphological Resources and Tools

A Case Study for Question-Answer Pairs

Delphine Bernhard
Bruno Cartoni
Delphine Tribout

Published by CSLI Publications

LiLT volume 5, issue 2 October 2011

A Task-Based Evaluation of French Morphological Resources and Tools

A Case Study for Question-Answer Pairs

Delphine Bernhard, *LIMSI-CNRS, Orsay (FR)* Bruno Cartoni, *LIMSI, CNRS, Orsay (FR), Dép. de Linguistique, Université de Genève* Delphine Tribout, *LIMSI-CNRS, Orsay (FR)*

Abstract

Morphology is a key component for many Language Technology applications. However, morphological relations, especially those relying on the derivation and compounding processes, are often addressed in a superficial manner. In this article, we focus on assessing the relevance of deep and motivated morphological knowledge in Natural Language Processing applications. We first describe an annotation experiment whose goal is to evaluate the role of morphology for one task, namely Question Answering (QA). We then highlight the kind of linguistic knowledge that is necessary for this particular task and propose a qualitative analysis of morphological phenomena in order to identify the morphological processes that are most relevant. Based on this study, we perform an intrinsic evaluation of existing tools and resources for French morphology, in order to quantify their coverage. Our conclusions provide helpful insights for using and building appropriate morphological resources and tools that could have a significant impact on the

LiLT Volume 5, Issue 2, October 2011.
A Task-Based Evaluation of French Morphological Resources and Tools.

application performance.

1 Introduction

Natural Language Processing (NLP) systems rely, in various extents, on language resources corresponding to the various levels of linguistic description, i.e. morphology, syntax or semantics. Among them, morphological knowledge is considered as a key resource for improving NLP applications. Inflectional morphology is broadly used, especially in POS-tagging tasks and for the rationalization of lexicon organization. However, when looking at the way derivational and compositional morphology is tackled in many NLP projects, it is striking to see how superficially morphological relations are sometimes addressed and how little attempt is made at using deep and motivated morphological knowledge. On the other hand, many research efforts and interesting findings have been made in the past decades in the understanding and automatic processing of word-formation processes, either in rule-based systems such as DériF (Namer, 2009) or empirical approaches like Linguistica (Goldsmith, 2001) or Morfessor (Creutz and Lagus, 2007). This lead to the development of tools of great coverage and complexity for the processing of morphology in various languages, and to the building of extended lexical resources for specific phenomena.

A first step towards bringing together recent advances in computational morphology and applied NLP projects is to assess the relevance of morphological knowledge in NLP applications, and to prioritize the morphological processes that have to be tackled. Surprisingly, few detailed evaluations have been carried out so far on the precise impact of morphological knowledge in NLP applications, and, to our knowledge, the rare published studies focus on the global improvement of systems. In this article, we describe an annotation experiment whose goal is to assess the contribution of morphological knowledge for one specific task, namely Question Answering (QA). This specific task has been chosen as being representative of many issues in other NLP applications such as Information Retrieval, reformulation and paraphrasing, Question Generation and Intelligent Tutoring Systems.

Based on a large set of manually annotated French question-answer pairs, we highlight the kind of linguistic knowledge that is necessary for this particular task and we propose a qualitative analysis of morphological phenomena to identify the morphological processes that are most relevant. Moreover, we perform an intrinsic evaluation of existing tools and resources for French morphology, in order to quantify their coverage. Finally, we argue that these kinds of studies are a prerequi-

site before building appropriate lexical resources that would have a significant impact on the application performance.

The article is organized as follows. Sections 2 and 3 respectively detail the field of morphology from the point of view of linguistics and Natural Language Processing. We introduce the research questions addressed in this article in Section 4. Section 5 details our annotation study of question-answer corpora. In Section 6 we report the quantitative and qualitative results of our annotation study. Finally, we assess the coverage of existing French morphological resources and tools for the morphological processes identified in Section 7 and conclude in Section 8.

2 Morphology in Linguistics

Morphology is the branch of linguistics that deals with words, their internal structure and the way they are formed. Classically, two distinct phenomena are considered: inflectional morphology and constructional morphology.[1] Inflectional morphology deals with the different forms a lexeme may take according to grammatical information such as gender, number, case, tense or mood. The different forms of a lexeme vary according to its syntactical context, and inflectional morphology is therefore often called *morphosyntax*. Constructional morphology (also called *morphosemantics*) deals with the formation of new lexemes by means of affixation, conversion or compounding. In these processes, the complex lexeme has a morphological and a semantic relation with its base lexeme.

Morphological processes may involve autonomous lexemes or suppletive forms.[2] A typical derivational process can apply on a base lexeme (e.g. *adapt*) to create a complex lexeme (e.g *adaptable*). In this particular example, the process forms an adjective out of a verb with the meaning 'that can be adapted', and appends the suffix *able*. Sometimes, a suppletive form of the base is used to coin complex lexemes (in French, the suppletive form of *jeu* 'play', *lud*, is used to coin the relational adjective *ludique*). Another frequent derivational process is *conversion*, where no specific marker (except for inflectional marks) is used to coin a new lexeme in a different category (like in the French *travailler* > *travail*, eng: *to work* > *work*).

For compounding, two (or more) base lexemes are used to coin a com-

[1] Even if there are sometimes no clear-cut boundaries between inflectional and constructional morphology.

[2] In this paper, we adopt a lexeme based approach to morphology, which denies the existence of morphemes, and gives the lexeme a central role

plex lexeme. In this case, the meaning of the complex lexeme is made of the two meanings of the two base lexemes, modulo some restrictions depending on the kind of construction as well as some pragmatic constraints, e.g. *firewood* is 'wood used to make fire' (Haspelmath, 2002).

In specialized language like medical terminology, a very frequent morphological process is neoclassical compounding (Cottez, 1984). This particular type of compounds is made of one (or more) so-called *neoclassical elements*, that are non-autonomous material, like affixes, but that differ from affixes because (i) they are from Latin or Greek origin, (ii) they are semantically related to an autonomous lexeme (and therefore still have a denotative/ontological meaning), and (iii) they can often be used indifferently at the beginning or at the end of the base-lexeme (*hydrothérapie* is coined with the neoclassical compound *hydro* that stands for *water* to mean "therapy with water").

Like other levels of linguistic description, morphology is addressed and used in many computational applications or tasks, as described in the following section.

3 Morphology in Natural Language Processing

In this section, we describe the various approaches to morphological analysis in NLP, and detail tools and resources as well as their use in various language technology applications.

3.1 Tools and Resources

Morphological analysis tools can be decomposed into rule-based systems and techniques relying on supervised or unsupervised learning.

Rule-based Systems

There are various types of morphological analysis methods based on rules: (i) simple methods, relying on heuristic rules, such as stemming algorithms, (ii) transducers, which are mainly used for the analysis and generation of inflectional morphology and (iii) morphosemantic analyzers, which perform both a semantic and a formal decomposition of lexemes

Stemming is often used in information retrieval. It consists in removing and transforming word endings so that words belonging to the same morphological family share an identical stem (Porter, 1980). These operations are encoded as language-specific heuristic rules. While stemming is rather efficient, especially for English, it nevertheless leads to somehow unnatural results, as stems do not necessarily correspond

to real words in the language considered.

Unlike stemming, *two-level morphology* (Koskenniemi, 1984) is non-directional and can be used both for analysis and generation. Rules are represented as transducers matching the surface level and the lexical level of analysis. Two-level morphology is particularly well suited for languages which make heavy use of the suffixation process like Finnish or Turkish (ten Hacken and Lüdeling, 2002). It is also able to deal with derivational morphology and compounding in order to derive an analysis of constructed words (Antworth, 1995).

Morphosemantic approaches give prominent weight to semantic information. Such systems have been proposed for the analysis of medical vocabulary in English (Pratt and Pacak, 1969), German (Hahn et al., 2003) and French (Lovis et al., 1995). The DériF system (Namer, 2009), originally developed for the French language has been successfully extended to the English medical vocabulary in order to account for neoclassical compounds (Deléger et al., 2009a). Cartoni (2009) presents a morphosemantic approach for the multilingual analysis of constructed neologisms in a source language and generation of their translation into a target language. However, morphosemantic systems usually suffer from lack of coverage, as they rely on detailed linguistic analyses which are often restricted to limited linguistic phenomena.

Supervised and Unsupervised Learning

Supervised and unsupervised learning methods aim at circumventing the need for manually developed rules by automatically acquiring morphological knowledge from either annotated or raw data.

Supervised learning approaches necessitate annotated training data. Both van den Bosch and Daelemans (1999) and Stroppa and Yvon (2006) use the CELEX database (Baayen et al., 1995) as a training dataset to learn morphological analyses using either memory-based learning (van den Bosch and Daelemans, 1999) or analogical learning (Stroppa and Yvon, 2006).

Annotated training datasets are not readily available for all languages. Moreover, they have to be constituted manually by experts, which costs both money and time. As a consequence, lots of research has been devoted during the last ten years to the development of methods for the unsupervised acquisition of morphological knowledge. These methods do not rely on language-specific resources and proceed from simple word lists or textual corpora. They usually aim at splitting words into morphemic segments or discover families of morphologically related words. Research in this domain has been recently fostered by

the Morpho Challenge, an international competition whose goal is to compare unsupervised approaches to morpheme discovery.[3]

Unsupervised morphology learning systems rely on a variety of cues and heuristics: graphical relatedness between words (Jacquemin, 1997, Gaussier, 1999, Zweigenbaum and Grabar, 2000), transition probability between characters (Harris, 1955, Keshava and Pitler, 2006, Bernhard, 2006, Spiegler et al., 2009), the MDL (Minimum Description Length) principle (Goldsmith, 2001, Creutz and Lagus, 2002) or advanced probabilistic models (Creutz and Lagus, 2005, Snyder and Barzilay, 2008, Spiegler et al., 2009).

As these methods usually encode only limited knowledge about morphology, their results are usually not as precise as systems relying on language-specific rules which have been manually developed.

Morphological Resources

In parallel to the tools and systems described above, another approach adopted to address morphological knowledge in NLP is the constitution of resources. While at the beginning, computer storage was an issue and dynamic approaches were preferred, constraint reductions in computer storage have made possible the constitution of large and specialized lexical databases. Compared to a dynamic processing of morphology (as described in the "tool" section above), resources are more static and rely greatly on a large amount of manual work. Moreover, these morphological resources are often phenomenon-specific and gather information about a specific lexical field, or specific morphological knowledge, e.g. derived nominals (MacLeod et al., 1997). Some automatic techniques, together with human work, have been used to build static lexical resources that provide morphosemantic information. For instance, Cartoni and Zweigenbaum (2010) describe an experiment to semi-automatically add inflectional information to a specialised medical lexicon for French by learning morphosyntactic information from existing lexicons.

If we consider more specifically the French language, which is the focus of this paper, inflectional morphology is well covered by several resources such as Morphalou,[4] Le*fff*[5] or Lexique 3.[6] However, constructional morphology is incompletely covered by several disparate resources (see Section 7.1 for a detailed description of these resources).

[3] http://www.cis.hut.fi/morphochallenge2009/
[4] http://www.cnrtl.fr/lexiques/morphalou/
[5] http://www.labri.fr/perso/clement/lefff/
[6] http://www.lexique.org/

Moreover, there is no large scale morphological resource for French, comparable to CELEX which has been built for Dutch, English and German (Baayen et al., 1995).

3.2 Applications of Morphology

Morphology is useful for a large variety of NLP applications, among which:[7]

- **Terminology**. Most of the work in terminological data processing focuses on the detection of lexical variants in complex terms, in order to fine-tune controlled term indexing and term extraction. Amongst the lexical phenomena that make a term vary in context, morphology (both inflectional and constructional) plays an important role (Jacquemin et al., 1997).
- **Speech Recognition and Synthesis**. Both applications can benefit from morphological information. Demberg et al. (2007) showed that morphological segmentations provided by rule-based systems improve the grapheme to phoneme conversion process. For speech recognition in morphologically complex languages such as Finnish, language models based on morphs instead of words are more robust with respect to the well-known problem of out of vocabulary words (Creutz et al., 2007).
- **Statistical Machine Translation**. Morphological analysis has been shown to improve the results in automatic translation between languages with different morphological structures (Lee, 2004).
- **Text Classification**. Linguistic units which are morphologically motivated can be used as features for text categorization, leading to a reduction of the dimension of the feature vectors and improved performance for a language with many compounds such as German (Witschel and Biemann, 2006).
- **Paraphrase Identification**. Paraphrases often contain morphologically related words, as for instance in the two following sentences: (1) *The new bridge was constructed by Wonderworks Ltd.* (2) *Wonderworks Ltd. is the constructor of the new bridge.*[8] Deléger and Zweigenbaum (2009, 2010) define paraphrase extraction rules targeted at medical documents for French and English relying on (i) nominalization and (ii) neo-classical compounding. They observe that nominal constructions are more frequent in technical texts written by experts (e.g., "treatment of the disease") while verbal con-

[7] For a complete overview of applications of morphology, see Daille et al. (2002).
[8] Example taken from Androutsopoulos and Malakasiotis (2010).

structions occur more frequently in documents targeted at lay readers (e.g., "the disease is treated").

- **Question Generation**. Question Generation systems usually apply syntactic transformation rules to the parse tree obtained for the input text (Gates, 2008, Heilman and Smith, 2009). Morphological information is mainly used for verb conjugation and is thus limited to inflectional morphology. One drawback of question generation methods is that the questions are usually very close to the text which served as a basis for their production. Constructional morphology could hence be used to automatically reformulate questions.

In this article, we analyze the kind of linguistic knowledge that is necessary for one specific task, namely Question Answering (QA), based on a large set of manually annotated French question-answer pairs. The choice of studying the role of morphology in QA applications was made because we believe they are representative of the typical issues raised by morphology in other NLP applications such as Information retrieval.

3.3 Morphology in Question Answering

QA systems aim at providing a precise answer to a given user question. To this aim, they usually rely on an Information Retrieval (IR) component which attempts to match words in the question and words in the text passages containing a potential answer. The major difficulty lies in the *lexical gap problem*, which occurs when a document is related to a question even though it does not contain the same words as the question. QA and IR systems must thus find a way of retrieving relevant documents without relying only on mere identity between words. In this context, morphology has often been preferred over semantics because the integration of morphological knowledge is easier. Research in IR and QA has thus tried to incorporate morphological knowledge, either by expanding the query, by indexing documents with morphologically motivated units or by using question reformulation or rephrasing patterns to identify the answer.

Most of the research carried out so far made use of simple heuristic-based stemming techniques which cut off word endings (such as (Lennon et al., 1988), (Harman, 1991), (Fuller and Zobel, 1998)). These turned out to be rather efficient for languages with a "less-rich" morphology, such as English, but they are not available for all languages (McNamee et al., 2009). In most cases, the recall is slightly improved, but these techniques also produce some noise, as shown by the example described in Bilotti et al. (2004): *organisation* and *organ* are stemmed to the same

form by the Porter Algorithm, even though synchronically there are not related any more. Another interesting piece of research, described in Moreau and Claveau (2006), shows that extending the query by morphological knowledge significantly improves the results, in most of the European languages for which they performed the experiment. To acquire morphological knowledge, they made use of a learning method based on analogy techniques. Consequently, they captured only affixation processes, and moreover only transparent affixation processes (that share a rather long character string), leaving aside conversion, reduction processes, or affixation on suppletive forms. They also admitted that, even with some precautions (long minimal character string, etc.), some incorrect pairs of morphologically related words are captured (*pondre* with *répondre*).

In each study mentioned so far, a word is conflated with all the lexemes belonging to its family according to the specific morphological analysis tool in use, since most of the tools do not distinguish between morphological phenomena, e.g. nominalization or relational adjectives. As a consequence, all words belonging to the same morphological family are considered as having approximately the same meaning. This may, in some cases, have a detrimental effect, if the morphological family induced is too broad. To our knowledge, no study has been carried out on deepest, rule-based, morphologically motivated methods for query expansion or document indexation.

Finally, a last method for integrating morphology in QA has to be mentioned, which consists in generating reformulations of the question or the answer, using morphology (Jacquemin, 2010), lexical semantics and textual inferences (Lin and Pantel, 2001, Hermjakob et al., 2002, Ravichandran and Hovy, 2002). Jacquemin (2010) describes a derivational resource for the French language which is automatically built from Dubois' dictionary of French verbs (Dubois and Dubois-Charlier, 1997). The rephrasing method first applies a form of word sense disambiguation based on verb senses in order to distinguish semantically motivated subgroups of derived words for a given term in context. The syntactic dependency structure of the candidate answer text is subsequently transformed by syntactic derivation patterns which integrate the derived terms. In contrast to this method, others aim at generating more general reformulation patterns, which are not limited to morphology but also rely on lexical semantics and textual inference. Lin and Pantel (2001) describe an unsupervised method to automatically acquire so called "inference rules" based on dependency trees. While inference rules encompass a large variety of phenomena, they also include purely morphological processes such as nominalization, e.g. "X

manufactures Y" ⇔ "X is manufacturer of Y". In closely related work, Hermjakob et al. (2002) and Ravichandran and Hovy (2002) detail patterns to derive reformulations of the question. Again, while these transformation patterns cover a wide range of phenomena, they also include a subset of morphological processes, e.g. "How deep is Crater Lake?" ⇒ "Crater Lake has a depth of ..." However, while reformulation methods largely rely on morphology, the focus generally lies in semantics and inference. Morphology has therefore not been studied in detail and is generally included in syntactic phenomena.

4 Research Questions

As we have shown, QA applications mostly rely on partial or superficial morphological knowledge. Moreover, only few studies specifically address the role of morphology within such systems. Most of the evaluations are extrinsic (based on the measurement of the improvement of an entire system when a morphological "module" is applied), and globally, the use of morphology (either indexing or query expansion) is very coarse.

However, some morphological analyzers and resources are now able to provide knowledge about a large spectrum of morphological processes. The issue is more in choosing and weighting the relations to be developed and implemented and in determining the resources or tools to be used or built if lacking. We hence address two specific research questions in this article:

- What morphological phenomena are most relevant in QA applications?
- How well do available tools and resources for French morphology cover these phenomena?

These two aspects are linked together because we first need to characterize the morphological relations which are relevant in QA in order to evaluate the use of resources and tools, and prioritizing the further constitution of resources.

To evaluate the impact of morphological knowledge on QA systems, we manually analyzed pairs made of a question on one side, and snippet containing the answer on the other side. These pairs come from three different corpora, as described in Section 5.1. First, for each pair we annotated words that are morphologically related on both sides. While doing so we also qualified the nature of this morphological relation as being inflectional, derivational or compositional. At the end of this annotation step, we obtained a list of pairs of morphologically related

words. We then analyzed these pairs according to different criteria: the part-of-speech involved, the grammatical features of the words in the case of inflectional relations, the kind of morphological process in the case of derivational relations, or the location of the most complex word (in the question or in the snippet).

We subsequently used this set of pairs of morphologically related words as a "gold-standard" to evaluate the coverage of available tools and resources for French. Since the gold-standard has been carefully characterized, precise measures can be computed for different morphological processes.

The contributions of the paper are as follows:

- We present the constitution and the analysis of a unique gold-standard for morphological relations, based on a detailed annotation of three different corpora of question-answer pairs. This study provides important insights on the type of morphological knowledge to be integrated into QA systems in order to improve their performance.
- We evaluate and compare two tools and five resources for French morphology, including both inflectional and derivational morphology.
- We make concrete proposals about the resources which would be most helpful for QA.

Section 5 below describes the annotation of the pairs, and Section 6 gives the results of the characterization of morphological processes in the set of pairs. Section 7 presents the task-oriented evaluation of several morphological resources and tools.

5 Annotation of Question-Answer Pairs

5.1 Description of the Datasets

The datasets gathered for the annotation come from three very different QA corpora, Quæro, EQueR-Medical and Conique, which are presented below. Our aim in annotating different types of corpora was to determine if there are significant differences in the morphological processes observed depending on the type of data. Table 1 presents statistical information on each corpus.

Quæro

The French Quæro corpus has been built for QA evaluation (Quintard et al., 2010) within the Quæro project. The corpus consists of 2.5M French documents extracted from the web and a set of 250 questions

	Conique	Quæro	EQueR-Medical
#Questions	201	350	200
#QA pairs	664	566	394
Avg. quest. length (words)	11.4	8.8	9.9
Avg. ans. length (words)	92.4	38.5	29.0

TABLE 1 Annotation corpora statistics.

for the 2008 evaluation and 507 questions for the 2009 evaluation. The document corpus has been constituted by taking the first 100 pages returned by the Exalead search-engine for a set of requests found in the search-engine's logs. As for the questions, they have been written by French native speakers by using the content of the documents for the 2008 evaluation, and by using only the requests' logs of the search-engine for the 2009 evaluation. There are three types of question: factual questions, boolean questions which ask for a yes-no answer and questions requiring a list for answer.

We constituted our corpus for the annotation task by taking all the snippets returned by the Ritel-QA System (Quintard et al., 2010) that have been manually validated as containing the correct answer for each factual question of the two evaluation campaigns. We thus obtained 566 pairs of question and snippet containing the answer, 338 from the the 2008 evaluation and 228 from the 2009 evaluation.

EQueR-Medical

The EQueR evaluation dataset has been constituted within the EQueR-EVALDA evaluation campaign for French Question Answering systems (Ayache et al., 2006). The campaign included two main tasks: (i) general domain QA over a collection of newspaper articles and senate reports and (ii) specialized domain QA over a collection of medical texts. We restricted our annotation study to the medical questions. The answer passages were retrieved by the participant systems and manually validated by a specialised judge.

Overall, the EQueR-Medical dataset comprises 394 question answer-passage pairs for 200 different questions.

Conique

The Conique corpus has been built with the objective of studying relevant answer justifications for QA systems (Grappy et al., 2010). Answer justifications provide additional material to the user, so that she/he may trust the answer retrieved by the system. The corpus is based on a subset of 291 questions from the French EQueR campaign (Ayache et al., 2006) and several CLEF campaigns. Candidate answer passages

Q: Quand **est né** Philippe d'Orléans? A: Philippe d'Orléans **naquit** le 2 août 1674.	**was born** **was born**
Q: Comment un *insuffisant* **rénal** doit-il être suivi? A: Du fait du risque de transmission nosocomiale du VHC chez les *insuffisants* **rénaux** hémodialysés et chez les transplantés rénaux, une surveillance annuelle de la sérologie doit être réalisée.	**CRF patient** **CRF patients**
Q: À combien de milliards de dollars s'élève le **déficit** budgétaire américain? A: Politique budgétaire. Le PIB des États-Unis s'élève à environ 10 000 milliards de dollars et les **déficits** atteindraient au moins 300 ou 400 milliards de dollars en 2003	**deficit** **deficits**

FIGURE 1 Q-A pairs involving an inflectional relation. The second column corresponds to a translation into English of the relevant word pair.

have been retrieved from the French Wikipedia using a coarse retrieval mechanism and manually annotated by seven annotators. In contrast to the two previously described datasets, answer passages in Conique do not correspond to the output provided by QA systems. It therefore constitutes an almost full recall dataset, devoid of any bias inherent to QA systems such as high question - answer passage token overlap.

We automatically pre-processed the annotated corpus to retrieve question - answer passages. We only kept full or partial justifications. Moreover, we reduced the passage to up to three sentences, centered on an annotated justification. Overall, the dataset we annotated comprises 664 question-answer pairs, for 201 different questions.

5.2 Annotation Methodology

The annotation was manually performed by three trained independent annotators,[9] using the YAWAT alignment tool (Germann, 2008). YAWAT was originally developed to align words in bilingual sentence-pairs for machine translation evaluation. In our case, we aligned and typed words and phrases in question-answer pairs. We defined three morphological tags: one for inflection, another for derivation and a third one for compounding. Figures 1-3 illustrate some question-answer pairs

[9] Co-authors of the present article.

involving inflectional, derivational and compounding relations.

Q: En quelle année Martin Luther King **a**-t-il **été assassiné**? A: Il dit avoir été à côté du pasteur King à Memphis lors de son **assassinat**, le 4 avril 1968.	**was assassinated** **assassination**
Q: Quels sont les quatre **réalisateurs** du film "Le jour le plus long"? A: Le Jour le plus long (The Longest Day) est un film américain **réalisé** par Ken Annakin, Andrew Marton, Bernhard Wicki et Gerd Oswald sorti en salle en 1962...	**directors** **directed**
Q: La pose d'amalgame dentaire peut-elle provoquer des **allergies**? A: Il est certain que la pose d'amalgames peut entraîner des réactions **allergiques** plus ou moins graves et prononcées chez les patients.	**allergies** **allergic**

FIGURE 2 Q-A pairs involving a derivational relation

Q: Où Marcos fut-il **dictateur**? A: Imelda Marcos, le 22 février 2006. Imelda Romualdez Marcos (née le 2 juillet 1929) fut la femme de Ferdinand Marcos, **président-dictateur** des Philippines de 1965 à 1986.	**dictator** **dictator president**
Q: Le mercure est-il un métal **toxique**? A: En grande concentration ou lorsque l'exposition est prolongée, le mercure a des effets **neurotoxiques** connus, principalement dans sa forme organique, soit le méthylmercure	**toxic** **neurotoxic**
Q: Qu'engendre la corticothérapie sur l'**os**? A: Les fractures de l'**ostéoporose** cortisonique surviennent au moins en partie en raison d'une perte osseuse induite par la corticothérapie	**bone** **osteoporosis**

FIGURE 3 Q-A pairs involving a compounding relation

Q: Quelle est la capitale de l' **Australie** ?
A: le territoire sur lequel est située la capitale fédérale **australienne**, Canberra .

FIGURE 4 Example of QA pair where both derivational and inflectional information are available

Since there can be more than one morphological step between two morphologically related words we defined specific guidelines for the annotation. First, we did not annotate inflectional variants of auxiliaries and determiners, as these tend to be very frequent but do not provide any interesting semantic information. Second, derivation and compounding supersede inflection. For instance, in the Q-A pair presented in Figure 4 there are two morphological steps between the noun *Australie* (eng: "Australia") in the question and the feminine adjective *australienne* (eng: "australian") in the answer: the first step is the derivation of the adjective *australien* (eng: "australian") out of the noun, and the second one is the inflection of the derived adjective in a feminine form. But the relevant morphological relation between the question and the answer is the derivation of the adjective *australien* out of the noun *Australie*, so that only this one has been annotated. Finally, a specific tag "other" was used to label words that are not directly related (i.e. that are related by more than one morphological process). In these cases, both members of the pairs are complex lexemes. In the characterization step (see section 6), these specific pairs have been treated differently.

5.3 Inter-annotator Agreement

Annotating morphological relations in question-answer pairs is not as easy a task as it may seem. The three annotators, yet experts in French morphology, did not always agree in the annotation tasks, as shown in Table 2 that presents the number of morphological relations that have been found by one, two or three annotators in the three corpora. The table also details inter-annotator agreement computed using Fleiss' kappa (Fleiss, 1971).[10]

As shown for the three corpora, common annotations (by 2 or 3 annotators) are numerous for inflection, fewer for derivation, and an important discrepancy is found for the compounding process, especially in

[10] κ has been computed by considering whether annotators agree that the question-answer pair contains at least a pair of word related by the morphological relation under consideration.

		3 ann.	2 ann.	1 ann.	Total	κ
Conique	**Infl.**	53	68	47	168	0.674
	Deriv.	76	56	73	205	0.729
	Comp.	1	20	13	34	0.39
Quæro	**Infl.**	71	42	23	136	0.83
	Deriv.	27	27	30	84	0.662
	Comp.	2	0	3	5	0.665
EQueR	**Infl.**	34	19	18	71	0.787
	Deriv.	30	38	33	101	0.662
	Comp.	10	26	47	83	0.442

TABLE 2 Number of identical annotations assigned by 1, 2 or 3 annotators and κ statistics for the three question-answer corpora.

the very specialized corpus EQueR. The κ varies accordingly from fair and moderate agreement for compounding and substantial to almost perfect agreement for inflection. These results highlight the difficulties in finding morphological relations, and classifying them into the three different classes (inflection, derivation and compounding).

However, when discussing the results between annotators, real disagreements were few and were mostly due to oversights. They also pointed at some limitations in the definitions of morphological phenomena when confronted with real data. Since the main goal of the annotation task was to create a gold-standard of morphological relations for the given task, disagreements have been resolved and a common set of data have been set up for the rest of the analysis. Details on the common set of data which the analyses are based on are found in Figure 5.[11]

6 Analysis of the Annotated Data

At the end of the annotation step, we obtained a set of morphologically related words, that can be studied according to different points of view. First we studied the repartition of morphological relation types such as inflection, derivation and compounding in the three corpora. Then, we analyzed in more details the part-of-speech involved in each morphological relation, the grammatical features expressed by the inflectional processes, the semantic types of derivational processes, and the location of the more complex word (in the question or in the answer).

[11] As disagreements have been resolved for the gold-standard dataset, the figures displayed in Figure 5 do not correspond to totals in Table 2.

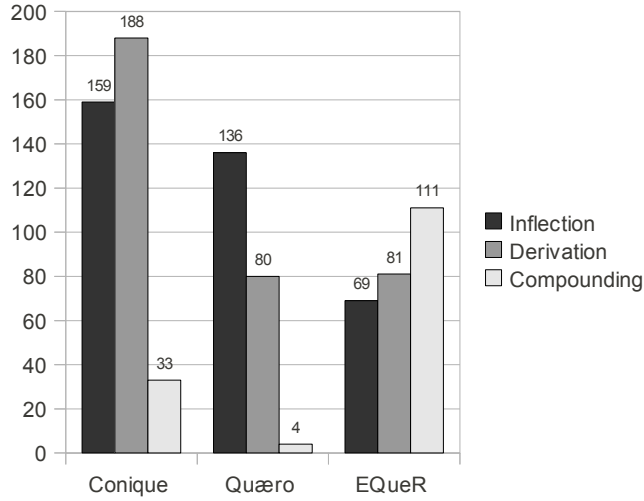

FIGURE 5 Inflection, derivation and compounding in the three corpora

6.1 Morphological Relation Types

The results of the annotation of each corpus according to the different types of morphological relations are presented in Figure 5. Each Q-A pair does not necessarily contain a morphological relation, and, more importantly, several pairs of words in the same question-answer pair can be morphologically related to one another, with different morphological relations.

Figure 5 shows that each corpus seems to favour a particular type of morphological relation: the Conique corpus contains a majority of derivational relations, while the Quæro corpus comprises more inflectional morphology. As for the EQueR corpus, it presents more compounding than any other kind of morphological relation. Moreover, if we consider the type of morphological relation depending on the corpus, inflection has the greatest proportion in the Quæro corpus, derivation is proportionally more present in the Conique corpus, while compounding is almost absent from Conique and Quæro and very important in EQueR.

The Conique and Quæro corpora show little difference with respect to the proportion of compounding. However, Conique contains more derivational relations than Quæro does. This is due to the way the Conique corpus has been built. It is not based on the output of a QA system, but the answers have been manually retrieved and annotated.

Q: Quelle est la conséquence de la **corticothérapie** sur l' os ?
A: Le problème essentiel des **corticoïdes** réside dans leurs effets secondaires (... ostéoporose, ostéonécrose aseptique des têtes fémorales ou parfois humérales ...).

FIGURE 6 Example Q-A pair from EQueR

QA systems usually have difficulties in dealing with derivational morphology. Moreover, there is a large variation in question and answer length between both corpora, as shown in Table 1. This could also explain the presence of more derivationally related pairs of words in Conique, because the longer the questions and the answers, the more opportunities to observe a derived word and its base. As for EQueR, the great proportion of compounds is certainly related to domain of the corpus: it contains a lot of medical terms, which are often compound nouns, as shown in Figure 6.

6.2 Inflection

The analysis of the inflectional relations found in the three corpora confirms the difference, already observed at the relation type level (Section 6.1), between Conique and Quæro on the one hand and EQueR on the other hand. Indeed, in Conique and Quæro most of the inflectional relations are verbal, whereas in EQueR most of them are nominal and the verbal ones are very few, as shown in Figure 7.

6.3 Derivation

As shown in Figure 5, derivation is important in the three corpora (between 30% and 50% of the pairs). In some cases the word in the question and the word in the answer are morphologically related by more than one derivational step. For instance *lune* (eng: "moon") and *alunissage* (eng: "landing on the moon") or *lait* (eng: "milk") and *allaitement* (eng: "breast-feeding"). In these cases one word is more complex than the other, but the complex word is not directly derived from the less complex one. In some other cases, like *joueur* (eng: "player") and *jouable* (eng: "playable") the word in the question and the word in the answer are morphologically related but neither derives from the other. Instead, they both derive from another word, which is *jouer* (eng: "play") for *joueur* and *jouable*. Table 3 shows the proportion of direct derivational relations, non direct derivational relations and cases where both words are complex and derived from another word. The figures show that

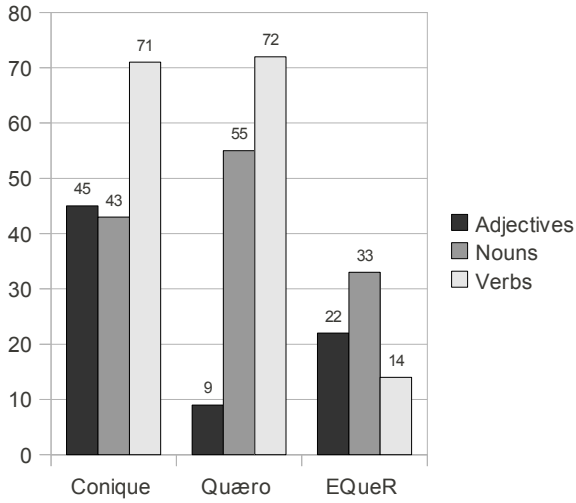

FIGURE 7 Parts of speech involved in inflectional processes

most of derivational relations between a word in the question and a word in the answer are direct (between 86% and 92%). Only very few relations are non direct.

	direct relation		**two steps**		**both complex**	
	#	%	#	%	#	%
Conique (188)	174	92.5	2	1.1	12	6.4
Quæro (80)	70	87.5	1	1.3	9	11.2
EQueR (81)	70	86.4	3	3.7	8	9.9

TABLE 3 Derivational steps between the word in the question and the word in the answer

There is little to say about the case where the derivational relation is non direct, since in that case the relation between the two words is difficult to predict. However, these cases would have an influence on the choice for the implementation methods, as we explain in Section 6.5. In this analysis step, we thus focus on the pairs which contain one base and one derivative, with only one derivational process between the two.

While focusing on the direct derivational relations, we can evaluate the proportion of different derivational processes involved. Table 4 presents the result of this evaluation. The figures in Table 4 show that the corpora differ with respect to the derivational processes used. While

Conique shows more denominal adjectives (about 50% of the derivational processes), Quæro and EQueR seem to favour noun formation processes (with respectively 61% and 54% of the derivational processes). These 2 particular derivational processes are described below.

	Conique (174)		**Quæro (70)**		**EQueR (70)**	
	#	%	#	%	#	%
Noun > Adj	41	23.6	16	22.9	28	40.0
Proper N > Adj	45	25.9	8	11.4	1	1.4
Noun > Noun	29	16.7	5	7.1	7	10.0
Proper N > N	6	3.4	8	11.4	2	2.9
Adj > Noun	3	1.7	0	0	4	5.7
Verb > Noun	41	23.6	30	42.9	25	35.7
Other	9	5.1	3	4.3	3	4.3

TABLE 4 Derivational processes in question-answer pairs

Denominal adjectives

In our data, all adjectives deriving from a proper noun (Proper N) are relational adjectives, like *chilien* (eng: "chilean") derived from *Chili* (eng: "Chile"), or *africain* (eng: "african") derived from *Afrique* (eng: "Africa"). Adjectives deriving from a common noun are mostly relational adjectives too, as shown by Figure 8. For instance *présidentiel* (eng: "presidential") derived from *président* (eng: "president"), or *solaire* (eng: "solar") derived from *soleil* (eng: "sun"). However there also are some adjectives with a predicative reading. For instance *âgé* (eng: "old") which derives from *âge* (eng: "age") with the meaning 'having a certain age' or *montagneux* (eng: "mountainous") derived from *montagne* (eng: "mountain") with the meaning 'full of mountains'. Figure 8 presents the proportion of adjectives (denominal or qualifying) in our corpora, and shows that denominal adjectives are much more frequent in the three corpora.

Noun Formation Processes

As regards the noun formation processes, the three corpora favor deverbal nominalizations, but deadjectival and denominal nominalizations are also found.[12] The formations of a noun out of a noun are very few, except in Conique. Those are mostly masculine and feminine profes-

[12]The type of nominalization (deverbal, deadjectival or denominal) depends on the part-of-speech category of the base (verb, adjective or noun, respectively).

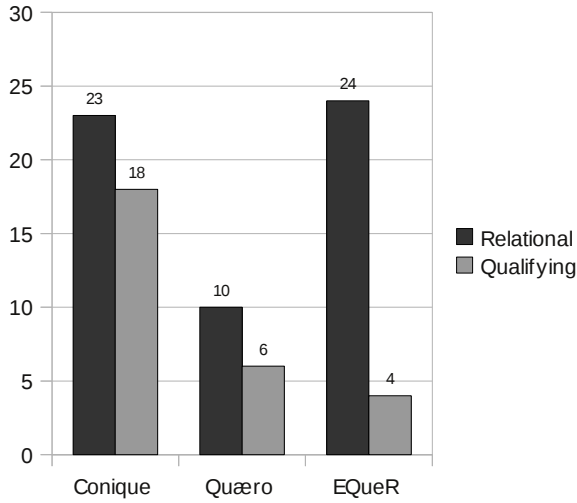

FIGURE 8 Types of denominal adjectives

sion names, like *infirmier* and *infirmière* (eng: "male/female nurse"), *directeur* and *directrice* (eng: "male/female director"), *président* and *présidente* (eng: "male/female president"), which we considered to be two distinct words rather than one and the same word inflected for gender. There are some suffixed diminutive nouns too, like *rame* (eng: "ream") > *ramette* (eng: "small ream"), and prefixed nouns like *président* (eng: "president") > *vice-président* (eng: "vice-president"). We also considered the formation of a noun out of a proper noun to be a denominal nominalization. These derived nouns are mostly demonyms which derive from a location denoting proper noun, like *Colombien* (eng: "Colombian") derived from the country name *Colombie* (eng: "Colombia"). This kind of nouns is found in the Conique and the Quæro corpora, but there are only two in the EQueR corpus, which is not surprising since it is a medical corpus.

Deadjectival nouns are very few in the three corpora. None of them is found in Quæro, and there are just a few of them in the other two corpora. These deadjectival nouns are property nouns, like *toxicité* (eng: "toxicity") which derives from the adjective *toxique* (eng: "toxic"). Not surprisingly deadjectival nouns denoting a property are mostly found in the EQueR corpus, since that corpus contains disease or trouble nouns, which often refer to the fact of being in a particular state.

As for deverbal nominalizations, they are most often event nouns in

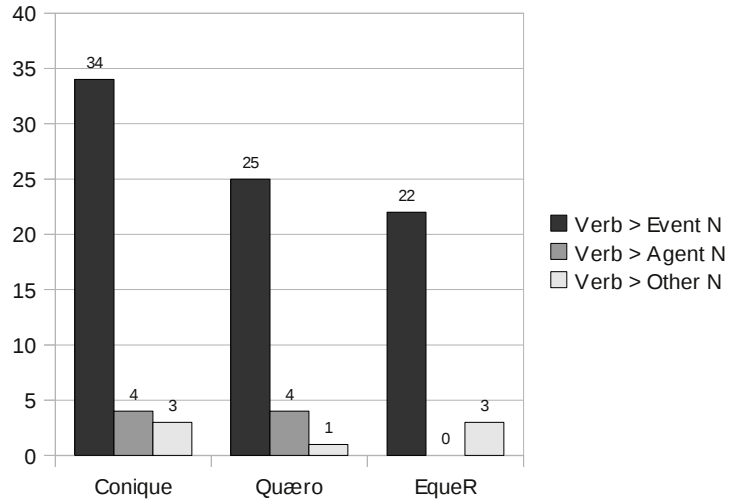

FIGURE 9 Semantic types of deverbal nouns in question-answer pairs

the three corpora, like *débarquement* (eng: "landing") derived from the verb *débarquer* (eng: "to land"). Event denoting nouns represent a large majority of the deverbal nouns, as shown in Figure 9. However, there is also a small number of agent nouns in the Conique and the Quæro corpora, but none in the EQueR corpus. For instance *réalisateur* (eng: "director") from *réaliser* (eng: "to direct"). And there is a small set of result nouns like *produit* (eng: "product") which derives from the verb *produire* (eng: "to produce").

Other Derivational Processes

Other derivational processes include for instance the formation of adverbs out of adjectives, like *complètement* (eng: "completely") derived from *complet* (eng: "complete"), or *directement* (eng: "directly") derived from *direct* (eng: "direct"). There also are some prefixed deverbal verbs like *déboucher* (eng: "unblock") out of *boucher* (eng: "block") or denominal adjectives like *international* (eng: "international") derived from *nation* (eng: "nation"). Interestingly we observed no deadjectival verb formation (like *national* "national" > *nationaliser* "nationalize") and almost no denominal verb formation. Only four denominal verbs were found in Conique, and none in the other corpora. Moreover, three of them are converted verbs: *border* (eng: "border") from *bord* (eng: "border"), *fusionner* (eng: "merge") from *fusion* (eng: "fusion") and *sui-*

cider (eng: "commit suicide") from *suicide* (eng: "suicide"). The only affixed denominal verb is *alunir* (eng: "land on the moon") derived from *lune* (eng: "moon"). The absence of denominal verbs could be explained by the rather unpredictable semantic relation between a noun and a derived verb. As stated by Hopper and Thompson (1984) there is an asymmetry in the lexical categories, since a nominalization still names the event denoted by the verb, whereas a verbalization does not refer to the entity denoted by the noun, but denotes an event associated with that entity. For instance, the noun *destruction* denotes the same event as its base verb *destruct*. But in the case of a denominal verb like *hospitalize*, the verb does not refer to the object denoted by the base noun *hospital*, but denotes some event related to that object. And the events we could associate to an entity are numerous and various. So, the semantic relation between a noun and its derived verb is less informative than that of a verb and its derived noun. It is not surprising then that so few nouns related to their derived verbs were found in the corpora.

Types of Morphological Operations: Affixation vs. Conversion

As regards the types of morphological processes, it is also worth noting that most of them are affixal. However, conversion as in *marcher* > *marche* (eng: *to walk* > *walk*) is also present in a more or less significant proportion depending on the corpus. In Conique and EQueR conversions represent about 10% of the derivational processes, while in Quæro the proportion of conversions rises up to 23%, as shown in Table 5.

	Affixation		**Conversion**	
	#	%	#	%
Conique (174)	153	87.9	21	12.1
Quæro (70)	54	77.1	16	22.9
EQueR (70)	63	90.0	7	10.0

TABLE 5 Proportion of affixation and conversion in derivational processes

We will show in section 7.2 that this distinction is important when assessing morphological tools, that deal differently with affixing or conversion processes.

6.4 Compounding

As regards compounds, we first analyzed whether the question-answer pairs contain a compound and at least one of its constituents (like *neurotoxique* and *toxique*, see Figure 3), or two compounds which share the same constituent (like *aéronautique* "aeronautics" and *aéroport* "airport" sharing the *aéro* constituent). Table 6 presents the result of this classification of compounds. The figures show that the three corpora have more pairs with a compound and its constituent(s) than pairs with two compounds. Moreover, pairs with two compounds are very few in Conique, and only EQueR has a significant number of such pairs.

	compound-constituent(s)		**2 compounds**	
	#	%	#	%
Conique (33)	26	78.8	7	21.2
Quæro (4)	4	100.0	0	0
EQueR (111)	70	63.1	41	36.9

TABLE 6 Cases where the pairs contain a compound and its constituent(s) and cases where the pairs contain two compounds sharing the same constituent

6.5 Intermediate Conclusion

As we have seen so far, morphology plays an important role in relating questions and their answers. More interestingly, we have seen that inflection is far from being the only kind of morphological knowledge involved in our corpus, and that both derivational and compositional relations play an important role.

These results show that processing morphological relations is important and we will show in Section 7 how existing resources and tools cover the needs. But before that, the set of morphological relations gathered in this study provides interesting insights to address strategical issues in implementing a QA system.

Location of the more Complex Word

Within all the morphologically related words gathered in this study, we assessed the location of the complex word. The predominance for one or the other position would influence importantly the way morphology is tackled in QA tasks. For derivational pairs, in the three corpora, the more complex word of a pair is most of the time in the answer, as shown in Table 7. For compounds, the results are very similar, as shown by

Table 8, apart from the EQueR corpus where compounds are more often in the question. This can be due to the type of questions. Indeed, in EQueR questions often are definitions of a medical term, like *Qu'est-ce qu'une aniridie ?* (eng: "What is a aniridia?"), or else questions about the symptoms or the treatment of a disease, and disease names often are complex terms, like *Citez 5 symptômes possibles d'une mastite* (eng: "Cite 5 possible symptoms of mastitis").

	Complex in Q		**Complex in A**		**2 Complex**	
	#	%	#	%	#	%
Conique (188)	51	27.1	**125**	**66.5**	12	6.4
Quæro (80)	24	30.0	**47**	**58.8**	9	11.2
EQueR (81)	31	38.3	**42**	**51.8**	8	9.9

TABLE 7 Position of the more complex word in a pair of derivationally related words

	Complex in Q		**Complex in A**		**2 Complex**	
	#	%	#	%	#	%
Conique (33)	3	9.1	**23**	**69.7**	7	21.2
Quæro (4)	1	25.0	**3**	**75.0**	0	0.0
EQueR (111)	**44**	**39.6**	26	23.5	41	36.9

TABLE 8 Position of the more complex word in a pair of words related by compounding

These results imply that in query expansion for QA, it is generally better to expand the query with the help of some derivatives of the question words, rather than analyse the words of the question and expand the query with their base words.

Direct or Indirect Morphological Relations

As shown in Section 6.3, and especially in Table 3, there are several morphologically related words that are not linked through a direct morphological process, but are either related through two morphological steps or are two complex words with the same word-base. Although this phenomenon is not very frequent (11.2 % for the most frequent cases: pairs of two complex words in the Quæro corpus), they also question the way morphology is tackled in specific tasks. Indeed, in such cases, query expansion would need to expand the query to all the words in the morphological family, which would be both time- and resource-consuming and also lead to a loss in precision.

However, the analyses performed on the data showed what kind of morphological process are most frequent (i.e. denominal adjective and noun formation processes), and these figures should help focusing on the most useful kind of expansion for the question or for the answer.

In the next section, we precisely evaluate the coverage of existing French resources with respect to these processes.

7 Task-oriented Evaluation of Morphological Resources and Analysers for French

7.1 Description of the Tools and Resources

Several resources and tools are available to deal with French morphology. In this article, we evaluate the coverage of five resources and two tools for French. Amongst these resources, Morphalou and Le*fff* address only inflectional knowledge while the three other, Verbaction, Dubois and Prolexbase, are specialized in morphosemantic phenomena. There are additional resources available for the French language, such as Morbocomp [13], Wiktionary[14], Polymots[15] or Morphonette[16]. However, a complete analysis of all resources for the French language is beyond the scope of this article whose main aim is to describe a task-based methodology for the evaluation of morphological tools and resources. As for the tools, one them, DériF, is specific to morphosemantics, and the other one, the Snowball stemmer, has been implemented to address both inflection and some construction phenomena. For a comparison of the DériF system and the Morphonette resource, we refer the reader to (Hathout and Namer, 2011).

Resources for inflection

Le*fff* is a syntactic and morphological lexicon for French (Sagot, 2010)[17]. It contains morpho-syntactic information for each inflected form, suchs as part of speech, lemma and sub-categorization. Over all it contains 534,763 inflected forms.

Morphalou is a freely available inflectional lexicon for French.[18] It contains 539,413 inflected forms corresponding to 68,075 lemmas.

There are also several inflection systems available for the French

[13] http://morbocomp.sslmit.unibo.it/
[14] http://fr.wiktionary.org/
[15] http://polymots.lif.univ-mrs.fr/v1/
[16] http://redac.univ-tlse2.fr/lexiques/morphonette.html
[17] http://alpage.inria.fr/~sagot/lefff.html
[18] http://www.cnrtl.fr/lexiques/morphalou/

language which are able to lemmatise words, among them TreeTagger (Schmid, 1994), FLEMM (Namer, 2000) and MElt (Pascal Denis and Benoît Sagot, 2009). Since most of these have already been thoroughly analysed and compared in a previous study by Namer (2000), we decided not to include them in this comparison and to focus on the recent large coverage resources Morphalou and Le*fff*. The latter is used in the MElt morphosyntactic analyser.

Resources for derivation

Verbaction is a lexicon of French action nouns linked to their corresponding verbs (Hathout et al., 2002, Hathout and Tanguy, 2002).[19] It totals 9,393 verb-noun pairs.

Dubois This resource is based on the description of French verbs by Dubois and Dubois-Charlier (1997).[20] It classifies verbs in semantic and syntactical classes and also provides information about adjectives and nouns derived from the verbs. Overall it contains 25,609 verb entries and mentions 33,955 derived words.

Prolexbase is a multilingual dictionary of proper nouns (Tran and Maurel, 2006, Bouchou and Maurel, 2008).[21] While not targeted at morphology, it nevertheless provides information about relational nouns and adjectives associated with proper nouns, e.g. *Français* and *français* (eng: "French") are explicitly associated with *France*. In some cases, relational nouns and adjectives are not morphologically related to the proper noun, e.g. *britannique* (eng: "british") with *Royaume-Uni* (eng: "United Kingdom"). Overall, it comprises 76,118 lemma and 20,614 derivational relations.

Tools

Among the two investigated tools, DériF is only designed for derivation and compounding while Snowball is more generic and can deal with both inflection and some derivational phenomena.

DériF (Namer, 2009) is a morphosemantic analyser for the French language, which is "based on decomposition rules and semantic interpretation templates" (Deléger et al., 2009b). It first has been designed for general language, and has also been adapted for medical specialised language. In this project, we use the online version, avail-

[19] `http://w3.erss.univ-tlse2.fr:8080/index.jsp?perso=hathout&subURL=verbaction/main.html`

[20] `http://rali.iro.umontreal.ca/Dubois/`

[21] `http://www.cnrtl.fr/lexiques/prolex/`

able at http://www.cnrtl.fr/outils/DeriF/. DériF analyses the lexeme given as an input (together with its morphosyntactic tag) and provides its base and its category. The process can be reiterated as long as the base found is morphologically complex. As an output, DériF also provides a "pseudo-definition" and a set of features describing the element used in the morphosemantic operation, as shown in Figure 10.

présidentiel/ADJ ==> [[présidence NOM] el ADJ]
"En rapport avec le(s) présidence"

FIGURE 10 Output of DériF for the relational adjective "présidentiel"

Snowball French Stemmer Stemming consists in removing and transforming word endings so that words belonging to the same morphological family share an identical stem (Porter, 1980). These operations are encoded as language-specific heuristic rules, which do not distinguish between inflection and construction. Moreover, the stem produced may not correspond to a word of the input language. The first stemmers have been developed for English, but versions for other languages are available. In our study, we used the Snowball French Stemmer.[22] Compared to the tools described before, the Snowball French Stemmer does not distinguish between inflection and construction. It is nevertheless limited to derivation.

7.2 Evaluation Results

One of the main objectives of this study is to measure the coverage of several existing lexical resources and tools for particular task (here QA) and describe in detail this task-based evaluation methodology. To perform this evaluation, the set of morphologically related words aligned by annotators is used as a gold-standard of morphological relations that need to be processed in this particular task. Thanks to this gold-standard, we can measure the coverage in percentage of pairs found in the resources or analyzed by the tool considered. The use of mixed resources and tools can also be assessed.

Inflection

As we described earlier, two resources (Morphalou and Le*fff*) and one tool (Snowball) are under consideration in this study to assess the pro-

[22] Freely available online at http://snowball.tartarus.org/download.php

Corpus (#)	Snowball		Morphalou		Le*fff*		Coverage	
	#	%	#	%	#	%	#	%
Conique (159)	140	88.0	157	98.7	159	100.0	159	100.0
Quæro (136)	105	77.2	135	99.3	135	99.3	136	100.0
EQueR (69)	65	94.2	65	94.2	65	94.2	68	98.5
Total (364)	**310**	**85.2**	**357**	**98.1**	**359**	**98.6**	**363**	**99.7**

TABLE 9 Coverage of inflection

cessing of inflectional phenomena. To evaluate the coverage of Morphalou and Le*fff*, each member of the pairs was checked against the lexicon. If correctly analysed, both members of the pairs would have the same lemma, and the link between them can be computed. For Snowball, the French version of the algorithm was run on the two members of the pairs, and the analysis is considered correct if a common minimal stem is found. Finally, global coverage was also calculated, by considering pairs that were correctly analysed by at least one of the resource/tool.

Table 9 presents the result of the evaluation of Snowball, Morphalou and Le*fff* for inflectional pairs. The figures show that Morphalou and Le*fff* are very similar and have a very good coverage. In lay corpora (Conique and Quæro), Morphalou and Le*fff* have a better coverage than Snowball. However in a specialized corpus such as EQueR, less words than in the other two corpora are found in Morphalou and Le*fff*, highlighting the specificity of the medical vocabulary. Snowball seems to be efficient enough for this kind of corpus and vocabulary. Moreover, the pairs of words which have been analyzed correctly by the three tools are not the same: Snowball is slightly better in analyzing adjectives and nouns, while Morphalou and Le*fff* succeed slightly better with verbs. Out of the total of the pairs (from the three corpora), Snowball appears to be less efficient: only 85% of the pairs are analyzed. But when used in parallel to Morphalou and Le*fff*, the global coverage reaches the very interesting level of 99.7%.

Derivation

Assessing derivational resources and tools is not as straightforward as inflection. Indeed, each tool and resource is designed to deal with lemmas only. Consequently, lemmatized forms of each member of the pairs were checked, taking for granted that lemmatization was performed successfully.

	Prolexbase	**VerbAction**	**Dubois**
N > A **proper N > A**	✓		
N > N **proper N > N** **A > N** **V > N**	✓	✓	✓

TABLE 10 Morphological phenomena addressed by derivational resources.

	VerbAction		**Dubois**	
Corpus (#)	#	%	#	%
Conique (34)	33	97.1	19	55.9
Quæro (25)	25	100.0	9	36.0
EQueR (22)	22	100.0	19	86.4
Total (81)	**80**	**98.8**	**47**	**58.0**

TABLE 11 Coverage of resources for deverbal event nouns

Morphological Resources for Derivation The three considered morphological resources that are available for French derivational morphology are designed to address specific morphological phenomena, which are summed up in Table 10: Dubois and VerbAction contain deverbal noun formation processes, while Prolexbase contains adjectives and noun deriving from proper nouns. More precisely, VerbAction covers deverbal event nouns only, while Dubois covers both agent and event nouns. As for Prolexbase, it covers demonyms and relational adjectives derived from geographical nouns. Consequently, assessing the relevance of these resources can only be done with the appropriate subpart of the gold-standard. The coverage of VerbAction and Prolexbase was calculated by counting the number of pairs that have been found in them. As for Dubois, it does not literally contain verbal derivatives. Those are only mentioned with specific information from which we can deduce the derivatives. Thus, in order to evaluate the coverage of Dubois we only took into account cases where the derivatives would be automatically computable from information provided in the resource.

As regards the deverbal nouns, Table 11 summarizes the coverage of VerbAction and Dubois for event nouns. As we can see, VerbAction has a better coverage than Dubois, especially in lay corpora (Conique and Quæro). As for the deverbal agent nouns, Dubois covers 100% of the Conique corpus and 75% of the Quæro corpus (no agentive noun

Corpus	Morphological relation (#)		Found in Prolexbase	
			#	%
Conique	Demonym - Rel Adj	(1)	1	100.0
	LocOrg - Demonym	(6)	6	100.0
	LocOrg - Rel Adj	(45)	43	95.6
Quæro	LocOrg - Demonym	(8)	5	62.5
	LocOrg - Rel Adj	(8)	8	100.0
EQueR	LocOrg - Rel Adj	(1)	1	100.0
Total		**69**	**64**	**92.7**

TABLE 12 Coverage of specific resources for geographic morphological relation

has been found in EQueR corpus), while VerbAction does not contain any of them, since it is devoted to action nouns. Dubois's coverage is narrower for this specific derivation process, but probably wider with respect to all the verb>noun derivational process.

As regards the demonyms and relational adjectives derived from geographical names, the result of the assessment of Prolexbase is presented in Table 12. We distinguished between Demonym (name for the resident of a place), Relational adjective, and LocOrg (grouping name of place and institutional entities). The figures show that Prolexbase has a very good coverage for both Demonyms derived from a Location name, and relational adjectives derived from Demonyms or Location names. In the Quæro corpus no Demonym>RelAdj pair has been found, and in the EQueR corpus, only one pair LocOrg>RelAdj has been found and is correctly analyzed in Prolexbase.

Morphological Tools for Derivation The two morphological analyzers DériF and Snowball are more generic than the derivational resources. Thus, they can be assessed on the entire gold-standard and not on specific sub-parts unlike resources. Table 13 presents the results of the evaluation of the two considered morphological analyzers. It also presents the global coverage of the two tools, as if used together.

Individually, both tools have a rather low coverage (under 50%) for every corpus, except DériF for the EQueR corpus. An interesting difference has to be noticed for the EQueR corpus, where DériF is much better than Snowball, and reaches 61% of analyses. This is due to the fact that DériF has been specifically extended to deal with medical terminology (Namer and Zweigenbaum, 2004).

Corpus (#)	Analysed by DériF #	Analysed by DériF %	Analysed by Snowball #	Analysed by Snowball %	Analysed by at least one tool #	Analysed by at least one tool %
Conique (174)	66	37.9	69	39.7	112	64.4
Quæro (70)	18	25.7	21	30.0	35	50.0
EQueR (70)	43	61.4	34	48.6	53	75.7
Total (314)	**127**	**40.4**	**124**	**39.5**	**200**	**63.7**

TABLE 13 Coverage of tools for derivational morphology

Corpus (#)		DériF #	DériF %	Snowball #	Snowball %
Conique (174)	affixed (153)	54	35.3	49	32.0
	converted (21)	12	57.1	20	95.2
Quæro (70)	affixed (54)	17	31.5	10	18.5
	converted (16)	1	6.2	11	68.7
EQueR (70)	affixed (63)	38	60.3	27	42.8
	converted (7)	5	71.4	7	100.0

TABLE 14 Coverage of tools for affixation and conversion

As for Conique and Quæro, it is surprising that Snowball gives better results than DériF. This can be explained if we consider the type of morphological operation involved. Indeed, both tools are not good at analyzing all affixation processes, but Snowball shows a larger coverage of conversion, as shown in Table 14. For this type of morphological operation it is not surprising that Snowball should perform better, since conversion usually corresponds to words with a large surface form overlap, if not identical. This makes a significant difference for the Quæro corpus where conversion is proportionally more important than in the other corpora (almost 23% of all derivational processes, against around 10% in Conique and EQueR; see Table 5 page 23 for the proportion of affixation and conversion in each corpus). The coverage of DériF is actually "affix-dependent" (not all formation processes have been implemented yet) and consequently it can be very good for some processes and far less appropriate for others.

It is also worth noting that the global coverage of both tools is between 50 % and 75% depending on the corpus, which shows that DériF and Snowball do not cover the same morphological relations.

Corpus (#)	Coverage of the resources		Coverage of the tools		Global Coverage	
	#	%	#	%	#	%
Conique (174)	96	55.2	112	64.4	149	85.6
Quæro (70)	41	58.6	35	50.0	58	82.9
EQueR (70)	26	37.1	53	75.7	59	84.3
Total (314)	**163**	**51.9**	**200**	**63.7**	**266**	**84.7**

TABLE 15 Coverage of all the tools and resources on derivational pairs

Global Coverage Finally, we evaluate the global coverage of resources and tools. For this specific evaluation, we count the number of pairs that have been analyzed at least by one of the resources or one of the tools, or both, which gives an interesting insight on the relevance of an ideal resource, mixing existing morphological knowledge and heuristic methods.

As can be seen in table 15, mixing all the three resources and the two tools proves to reach an interesting level of coverage. Indeed, the global coverage of resources and tools is higher than 80%. Morphological resources are efficient for specific morphological processes that are difficult to address with rule-based models (like the conversion process). Rule-based models, especially those that are linguistically-motivated like DériF, prove to be very efficient for specialized language, where complex morphological processes are involved.

However, frequent phenomena seem to be lacking in most of the tools considered, e.g., pairs involving relational adjectives and their base noun are not always covered, as shown in table 16.[23] This particular process is the second most frequent in the pairs in Conique and Quæro, and the most frequent in EQueR, as shown in section 6. Consequently, efforts in building resources or rule-based analyzers should be put on this particular phenomenon to address such a frequent issue, and increase importantly the coverage of such tools. This is also in some extent true for deverbal agent nouns which are lacking a dedicated and directly usable resource, even though the phenomenon is not as important as denominal adjectives.[24]

[23]These figures do not take demonymic adjectives into account

[24]Dubois does contain information about deverbal agent nouns. However, these nouns are not explicitly part of the resource and would have to be automatically computed from the indications provided in the resource.

Corpus (#)	noun-rel.adj. pairs	Analysed by DériF		Analysed by Snowball	
		#	%	#	%
Conique (174)	23	14	60.9	2	8.7
Quæro (70)	10	6	60.0	0	0.0
EQueR (70)	24	14	58.3	3	12.5
Total (314)	**57**	**34**	**59.6**	**5**	**8.8**

TABLE 16 Coverage of the tools for relational adjectives

Corpus (#)	Analyzed by DériF	
	#	%
Conique (26)	2	7.7
Quæro (4)	1	25.0
EQueR (70)	44	62.9
Total (100)	**47**	**47.0**

TABLE 17 Coverage of DériF for compounding morphology

Compounding

As for compounding, we have only evaluated DériF, because there is no specific resource dealing with compounding in French, and Snowball does not deal with compounding either. We provide statistics only for pairs that contain one (or two) constituent(s) on one side, and the compound on the other side, and we left apart other pairs, that were aligned because they contain two compounds, coined on a common constituent. In the gold-standard, a few annotated pairs were made of two derived lexemes, or of two compounds that have one constituent in common.

Table 17 presents the result of this evaluation of DériF. The figures show that compounding is a very complex process. DériF proves to be quite efficient for this particular process, even though still only less than half of the compounds are analyzed.

Quality of Analysis

Although this study focuses specifically on their coverage, the evaluation process highlights a few cases of wrong analyses made by morphological resources and tools.

Incorrect analyses appear especially for rule-based tools (such as DériF), whereas within resources, mistakes come from wrong encod-

ing by human annotators. Few errors were found in the analysis from DériF. In the Conique corpus, 11 pairs of related words were wrongly analyzed (5 in Quæro and 0 in EQueR). Errors mostly come from ambiguous character strings, that lead to a wrong decomposition by the tool. Most of the erroneous cases come from the application of a specific rule for specialized language on lexemes from the general vocabulary, e.g. the proper noun *Serbie* (eng: *Serbia*) is analyzed as an "infection" or "disease" related to *serbe*, because of the suffix *ie* which is typical of a disease in the medical domain, e.g. "thalassémie" (eng: *thalassemia*).

8 Conclusion and Perspectives

In this paper, we presented an in-depth analysis of the role of morphology in a specific NLP task: Question Answering. Based on a large-scale annotation of three distinct corpora of question-answer pairs, we built a gold-standard of morphologically related words in question-answer pairs. The gold-standard provides interesting insights on the position of the complex terms and on the kind of morphological relations that are often implied. Then, based on this set, we evaluated the coverage of existing resources and tools.
The main findings of this study are as follows:

1. The type of morphological processes involved vary depending on the type of the texts under consideration: compounding is very frequent in specialized medical texts but almost absent in general language.
2. Denominal adjectivations and deverbal nominalizations are the most frequent derivational phenomena encountered in our corpora. As a consequence, these phenomena should be addressed in priority when building a QA application.
3. Morphologically complex words are mostly found in the answer. Consequently, an appropriate strategy should be adopted for query expansion.
4. French derivational resources (either for demonyms or deverbal nouns) have a good coverage of the morphological knowledge they target.
5. However, some morphological phenomena are covered neither by the resources, nor by the tools that we analyzed, such as relational adjectives that are not demonymic.
6. Comparing a linguistically-motivated tool (DériF) and a heuristic tool (Snowball) shows the limits of the latter, but also the interesting complementarity of the two approaches.

In the future, we would like to integrate our findings in a specific NLP application, either question paraphrasing for Question Generation or reformulation patterns for QA. Moreover, statistics gathered in this study will help prioritizing which morphological processes need to be recorded in morphological resources or tools, and encourage the development of new French morphological resources for other phenomena.

Acknowledgments

This work has been partially financed by OSEO under the Quaero program.

References

Androutsopoulos, Ion and Prodromos Malakasiotis. 2010. A Survey of Paraphrasing and Textual Entailment Methods. *Journal of Artificial Intelligence Research* 38:135–187.

Antworth, Evan L. 1995. *User's Guide to PC-KIMMO Version 2*. `http://www.sil.org/pckimmo/v2/doc/guide.html`.

Ayache, Christelle, Brigitte Grau, and Anne Vilnat. 2006. EQueR: the French Evaluation campaign of Question-Answering Systems. In *Proceedings of LREC 2006*.

Baayen, R. Harald, Richard Piepenbrock, and Léon Gulikers. 1995. *The Celex Lexical Database (Release 2) [CD-ROM]*. Philadelphia, PA: Linguistic Data Consortium.

Bernhard, Delphine. 2006. Unsupervised Morphological Segmentation Based on Segment Predictability and Word Segments Alignment. In M. Kurimo, M. Creutz, and K. Lagus, eds., *Proceedings of the Pascal Challenges Workshop on the Unsupervised Segmentation of Words into Morphemes*, pages 19–23. Venice, Italy.

Bilotti, Matthew W., Boris Katz, and Jimmy Lin. 2004. What Works Better for Question Answering: Stemming or Morphological Query Expansion. In *Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR 2004*. Sheffield, England.

Bouchou, Béatrice and Denis Maurel. 2008. Prolexbase et LMF: vers un standard pour les ressources lexicales sur les noms propres. *Traitement Automatique des Langues* 49(1):61–88.

Cartoni, Bruno. 2009. Lexical Morphology in Machine Translation: A Feasibility Study. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 130–138. Athens, Greece.

Cartoni, Bruno and Pierre Zweigenbaum. 2010. Semi-Automated Extension of a Specialized Medical Lexicon for French. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, and D. Tapias,

eds., *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta.

Cottez, Henri. 1984. *Dictionnaire des structures du vocabulaire savant. Éléments et modèles de formation.*. Paris: Le Robert, 3rd edn.

Creutz, Mathias, Teemu Hirsimäki, Mikko Kurimo, Antti Puurula, Janne Pylkkönen, Vesa Siivola, Matti Varjokallio, Ebru Arisoy, Murat Saraçlar, and Andreas Stolcke. 2007. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Transactions on Speech and Language Processing (TSLP)* 5(1):1–29.

Creutz, Mathias and Krista Lagus. 2002. Unsupervised Discovery of Morphemes. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, pages 21–30.

Creutz, Mathias and Krista Lagus. 2005. Inducing the Morphological Lexicon of a Natural Language from Unannotated Text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, pages 106–113. Espoo, Finland.

Creutz, Mathias and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.* 4(1):3.

Daille, Béatrice, Cécile Fabre, and Pascale Sébillot. 2002. *Many Morphologies*, chap. Applications of Computational Morphology, pages 210–234. Cascadilla Press.

Deléger, Louise, Fiammetta Namer, and Pierre Zweigenbaum. 2009a. Morphosemantic parsing of medical compound words: Transferring a French analyzer to English. *International Journal of Medical Informatics* 78:48–55.

Deléger, Louise, Fiammetta Namer, and Pierre Zweigenbaum. 2009b. Morphosemantic parsing of medical compound words: Transferring a French analyzer to English. *International Journal of Medical Informatics* 78:48–55. MedInfo 2007.

Deléger, Louise and Pierre Zweigenbaum. 2009. Extracting Lay Paraphrases of Specialized Expressions from Monolingual Comparable Medical Corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, pages 2–10. Singapore.

Deléger, Louise and Pierre Zweigenbaum. 2010. Identifying Paraphrases between Technical and Lay Corpora. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, and D. Tapias, eds., *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta.

Demberg, Vera, Helmut Schmid, and Gregor Möhler. 2007. Phonological Constraints and Morphological Preprocessing for Grapheme-to-Phoneme Conversion. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 96–103. Prague, Czech Republic.

Dubois, Jean and Françoise Dubois-Charlier. 1997. *Les verbes français*. Larousse-Bordas.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5):378–382.

Fuller, M. and J. Zobel. 1998. Conflation-based comparison of stemming algorithms. In *Proceedings of the Third Australian Document Computing Symposium*, pages 8–13. Sydney.

Gates, Donna M. 2008. *Automatically Generating Reading Comprehension Look-Back Strategy Questions from Expository Texts*. Master's thesis, Carnegie Mellon University.

Gaussier, Eric. 1999. Unsupervised learning of derivational morphology from inflectional lexicons. In *Proceedings of the Workshop on Unsupervised Methods in Natural Language Processing*. University of Maryland.

Germann, Ulrich. 2008. Yawat: yet another word alignment tool. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (HLT '08)*, pages 20–23.

Goldsmith, John. 2001. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics* 27(2):153–198.

Grappy, Arnaud, Brigitte Grau, Olivier Ferret, Cyril Grouin, Véronique Moriceau, Isabelle Robba, Xavier Tannier, Anne Vilnat, and Vincent Barbier. 2010. A Corpus for Studying Full Answer Justification. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, eds., *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta.

Hahn, Udo, Martin Honeck, and Stefan Shulz. 2003. Subword-Based Text Retrieval. In *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)*. Big Island, Hawaii.

Harman, Donna. 1991. How effective is suffixing? *Journal of the American Society of Information Science* 42(1):7–15.

Harris, Zellig. 1955. From phoneme to morpheme. *Language* 31(2):190–222.

Haspelmath, Martin. 2002. *Understanding morphology*. Understanding Language. Londres: Arnold.

Hathout, Nabil and Fiammetta Namer. 2011. Règles et paradigmes en morphologie informatique lexématique. In *Actes de TALN 2011*. Montpellier, France.

Hathout, Nabil, Fiammetta Namer, and Georgette Dal. 2002. *Many Morphologies*, chap. An Experimental Constructional Database : The MorTAL Project, pages 178–209. Cascadilla Press.

Hathout, Nabil and Ludovic Tanguy. 2002. Webaffix: Discovering Morphological Links on the WWW. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1799–1804. Las Palmas de Gran Canaria, Espagne.

Heilman, Michael and Noah A. Smith. 2009. Question Generation via Overgenerating Transformations and Ranking. Technical Report CMU-LTI-09-013, Language Technologies Institute, Carnegie Mellon University.

Hermjakob, Ulf, Abdessamad Echihabi, and Daniel Marcu. 2002. Natural Language Based Reformulation Resource and Wide Exploitation for Question Answering. In *Proceedings of the Eleventh Text Retrieval Conference (TREC 2002)*.

Hopper, P.J. and S.A. Thompson. 1984. The discourse basis for lexical categories in universal grammar. *Language* 60:703–752.

Jacquemin, Bernard. 2010. A Derivational Rephrasing Experiment for Question Answering. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, eds., *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta.

Jacquemin, Christian. 1997. Guessing morphology from terms and corpora. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 156 – 165.

Jacquemin, Christian, Judith L. Klavans, and Evelyne Tzoukermann. 1997. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *Proceedings, 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL - EACL'97)*, pages 24–31.

Keshava, Samarth and Emily Pitler. 2006. A Simpler, Intuitive Approach to Morpheme Induction. In *Proceedings of the Pascal Challenges Workshop on the Unsupervised Segmentation of Words into Morphemes*, pages 31–35. Venice, Italy.

Koskenniemi, Kimmo. 1984. A general computational model for word-form recognition and production. In *Proceedings of the 22nd annual meeting on Association for Computational Linguistics*, pages 178–181.

Lee, Young-Suk. 2004. Morphological Analysis for Statistical Machine Translation. In D. M. Susan Dumais and S. Roukos, eds., *Proceedings of HLT-NAACL 2004: Short Papers*, pages 57–60.

Lennon, Martin, David S. Pierce, Brian D. Tarry, and Peter Willett. 1988. An evaluation of some conflation algorithms for information retrieval. *Journal of Information Science* 3(4):177–183.

Lin, Dekang and Patrick Pantel. 2001. Discovery of Inference Rules for Question Answering. *Natural Language Engineering* 7(4):343–360.

Lovis, Christian, Pierre-André Michel, Robert Baud, and Jean-Raoul Scherrer. 1995. Word Segmentation Processing: A Way to Exponentially Extend Medical Dictionaries. In R. A. Greenes, H. E. Peterson, and D. J. Protti, eds., *Proc 8th Word Congress on Medical Informatics*, pages 28–32.

MacLeod, Catherine, Adam Meyers, Ralph Grishman, Leslie Barrett, and Ruth Reeves. 1997. Designing a dictionary of derived nominals. In *Recent*

Advances in Natural Language Processing, Selected papers from RANLP '97, pages 45–56.

McNamee, Paul, Charles Nicholas, and James Mayfield. 2009. Addressing morphological variation in alphabetic languages. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 75–82.

Moreau, Fabienne and Vincent Claveau. 2006. Extension de requêtes par relations morphologiques acquises automatiquement. In *Actes de la Troisième Conférence en Recherche d'Informations et Applications CORIA 2006*, pages 181–192.

Namer, Fiammetta. 2000. FLEMM : un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues* 41(2):523–547.

Namer, Fiammetta. 2009. *Morphologie, Lexique et TAL : l'analyseur DériF*. TIC et Sciences cognitives. London: Hermes Sciences Publishing.

Namer, Fiammetta and Pierre Zweigenbaum. 2004. Acquiring meaning for french medical terminology: contribution of morphosemantics. *Eleventh MEDINFO International Conference* pages 535–539.

Pascal Denis and Benoît Sagot. 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Pacific Asia Conference on Language, Information and Computation*.

Porter, Martin F. 1980. An algorithm for suffix stripping. *Program* 14(3):130–137.

Pratt, Arnold W. and Milos G. Pacak. 1969. Automated processing of medical English. In *Proceedings of the 1969 conference on Computational linguistics*, pages 1–23.

Quintard, Ludovic, Olivier Galibert, Gilles Adda, Brigitte Grau, Dominique Laurent, Véronique Moriceau, Sophie Rosset, Xavier Tannier, and Anne Vilnat. 2010. Question Answering on Web Data: The QA Evaluation in Quæro. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, eds., *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta.

Ravichandran, Deepak and Eduard Hovy. 2002. Learning surface text patterns for a Question Answering system. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*, pages 41–47.

Sagot, Benoît. 2010. The Le*fff*, a freely available, accurate and large-coverage lexicon for French. In *Proceedings of LREC 2010*.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Snyder, Benjamin and Regina Barzilay. 2008. Unsupervised Multilingual Learning for Morphological Segmentation. In *Proceedings of ACL-08: HLT*, pages 737–745. Columbus, Ohio.

Spiegler, Sebastian, Bruno Golenia, and Peter Flach. 2009. PROMODES: A Probabilistic Generative Model for Word Decomposition. In *Working Notes for the CLEF 2009 Workshop*. Corfu, Greece.

Stroppa, Nicolas and François Yvon. 2006. Du quatrième de proportion comme principe inductif: une proposition et son application à l'apprentissage de la morphologie. *Traitement Automatique des Langues* 47(1):33–59.

ten Hacken, Pius and Anke Lüdeling. 2002. Word Formation in Computational Linguistics. In *Proceedings of TALN 2002*, vol. 2, pages 61–87.

Tran, Mickaël and Denis Maurel. 2006. Prolexbase : un dictionnaire relationnel multilingue de noms propres. *Traitement Automatique des Langues* 47(1):115–139.

van den Bosch, Antal and Walter Daelemans. 1999. Memory-based morphological analysis. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 285–292.

Witschel, Hans Friedrich and Chris Biemann. 2006. Rigorous dimensionality reduction through linguistically motivated feature selection for text categorization. In S. Werner, ed., *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, vol. 1 of *Ling@JoY : University of Joensuu electronic publications in linguistics and language technology*, pages 197–204.

Zweigenbaum, Pierre and Natalia Grabar. 2000. Liens morphologiques et structuration de terminologie. In *Actes de IC 2000 : Ingénierie des Connaissances*, pages 325–334.